# Ensemble Calibration of 500 hPa Geopotential Height and 850 hPa and 2-Meter Temperatures Using Reforecasts

Thomas M. Hamill and Jeffrey S. Whitaker

*NOAA Earth System Research Laboratory, Physical Sciences Division*
*Boulder, Colorado*

Submitted as a NOTE to Monthly Weather Review

8 November 2006

Corresponding author address:

Dr. Thomas M. Hamill
NOAA / ESRL, Physical Sciences Division
R / PSD 1, 325 Broadway
Boulder, CO 80305
Phone: (303) 497-3060
Fax: (303) 497-6449
E-mail: tom.hamill@noaa.gov

ABSTRACT

An examination of the benefits of ensemble forecast calibration was performed for 3 variables, 500-hPa geopotential height (Z500), 850-hPa temperature (T850), and 2-meter temperature (T2M). A large reforecast data set was used for the calibration. Two calibration methods were examined, a correction for a gross bias in the forecast, and an analog method that implicitly adjusted for bias, spread, and applied a downscaling where appropriate. The characteristics of probabilistic forecasts from the raw ensemble were also considered. Forecasts were evaluated using rank histograms and the continuous ranked probability skill score. T2M rank histograms showed high population of extreme ranks at all leads, and a correction for model bias alleviated this only slightly. The extreme ranks of Z500 rank histograms were slightly underpopulated at short leads, though slightly overpopulated at longer leads. T850 had characteristics in between those of T2M and Z500. Accordingly, Z500 was the most skillful variable without calibration and the variable least improved by calibration, and the bias correction achieved most of the improvement in skill. For T850, there was a more substantial additional increase in skill relative to the bias correction when the analog technique was applied. For T2M forecasts, probabilistic forecasts from the raw ensemble were the least skillful, the application of a bias correction substantially increased the skill, and the application of the analog technique produced the largest further increase in skill relative to the bias correction. Hence, reforecast data sets may be particularly helpful in the improvement of probabilistic forecasts of the variables that are most directly relevant to many forecast users, the sensible surface-weather variables.

## 1. Introduction

This note considers the effects of calibrating ensemble forecasts of three variables, 500-hPa geopotential height (Z500), 850-hPa geopotential temperature (T850), and 2-m temperature (T2M). Calibration here refers to the adjustment of the ensemble forecast probabilities to account for model bias, spread deficiencies, and/or a downscaling from the model grid to observation sites. Ideally, forecast probabilities would be reliable and sharp when calculated directly from the event frequency in the raw ensemble. However, typically the forecasts are contaminated by systematic biases or deficiencies in spread. Consequently, there is a growing body of literature now offering many possible ways of calibrating ensemble forecasts (Hamill and Colucci 1998; Eckel and Walters 1998; Roulston and Smith 2003; Wang and Bishop 2005, Raftery et al. 2005, Gneiting et al. 2005).

This article returns to consider calibration using *reforecasts*, a very large set of forecasts utilizing a stable model and data assimilation system. Previous articles (Hamill et al. 2006, Hamill and Whitaker 2006) have considered how to calibrate precipitation forecasts, which may require special techniques because the forecast probability density functions (pdfs) are typically not normally distributed. Here, the calibration of Z500, T850, and T2M appears to be simpler at first glance, for the forecast pdf and their errors may be approximately Gaussian. In such cases, a variety of calibration techniques may work well. Wilks and Hamill (2006) demonstrated that a number of parametric and nonparametric calibration techniques were suitable and several of the best were generally similar in their resultant skill for the problem of calibrating daily maximum and

minimum 2-meter temperature forecasts. Wilks (2006a) also provides a summary of many of the proposed calibration techniques and intercomparison using a simple model.

We intend here to address two primary questions related to the calibration of Z500, T850, and T2M: first, how much skill improvement is produced from a simple elimination of gross bias in the model forecasts, and how much more is added by the application of a technique that also accounts for spread deficiencies in the ensemble and produces an implicit statistical downscaling? Second, is calibration inherently easier or more difficult for one variable vs. another? Since the dawn of numerical weather prediction, forecasters have examined the properties of Z500. When considering calibration issues, is this a good proxy for understanding the characteristics of the more commonly utilized surface-weather variables?

This note will not discuss the calibration properties as a function of training sample size; the recent Wilks and Hamill (2006) manuscript discusses this in depth for temperature forecasts and Hamill et al. (2006) addresses this for precipitation forecasts. The manuscript also will not present a comparison of many of the proposed calibration techniques; for such a comparison, see Wilks (2006).

Section 2 below will describe the extensive reforecast data set used in this calibration experiment as well as the verification data and techniques. Section 3 provides results, and section 4 a brief conclusion.

## 2. Data set, methods of calibration, and verification techniques.

### *a. Data*

In this experiment, we explored methods of calibrating ensemble forecasts using reforecasts and observations. Our reforecast data set consisted of a 15-member ensemble forecast conducted from every day from 1979 to present, starting from 0000 UTC initial conditions. The forecasts extended to 15 days lead, with data archived every 12 h. The model was a 1998 version of the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS). It was spectral, with horizontal truncation at global wavenumber 62, with 28 vertical levels. The data was archived on a 2.5-degree grid. The ensemble initial conditions consisted of a control, initialized with the NCEP-National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996), plus a set of 7 bred pairs of initial conditions (Toth and Kalnay 1993, 1997) re-centered each day on the reanalysis initial condition. The breeding method was the same as that used operationally in January, 1998. For our purposes, we considered only the calibration of 500 hPa geopotential, 850-hPa temperature, and 2-m temperature, and we processed forecasts from 1979 to 2004, a total of 26 years of forecasts. The reforecast data set and model characteristics are described more completely in Hamill et al. (2006), which also provides a link for users to download the data set themselves.

NCEP-NCAR reanalysis data (Kalnay et al. 1996) was used for verification of 500-hPa heights (hereafter, Z500) and 850 hPa temperatures (hereafter, T850), and calibration for these variables was performed over the Northern Hemisphere. 0000 UTC surface observations over North America were used to demonstrate the calibration of 2-m temperature (hereafter, T2M). We limited our calibration to stations where observations were available for 96 percent or more of the days between 1979 and 2004, a total of 292 stations (Fig. 1). The source of these surface observations was the National Center for

Atmospheric Research's (NCAR's) data set DS472.0. While it would have been preferable for comparison purposes to calibrate against a gridded field, the 2-m temperature fields from the NCEP-NCAR reanalysis are not particularly useful; the analysis method does not utilize near-surface observations and inherits systematic errors from the first-guess fields. We are unaware of any other high-quality, unbiased, gridded temperature reanalyses.

*b. Calibration methods.*

1) ENSEMBLE RELATIVE FREQUENCY ("RAW")

The simplest approach used no statistical calibration. The relative frequency of event occurrence was estimated directly from the 15-member ensemble.

2) "BIAS-CORRECTED" RELATIVE FREQUENCY

In this procedure, probabilistic forecasts were generated from an ensemble of forecasts, where each member has been bias-corrected according to the long-term bias statistics in a cross-validated manner (Wilks 2006b, p. 215). Bias corrections were calculated separately for each grid point, each year, and each day and were based on a centered, 31-day running-mean difference between the forecast and observed climatology. For example, when calibrating a 1 February 1979 forecast, the forecasts and observations excluded 1979 data but included data from +/- 15 days around 1 February and data from 1980-2004, a total of 31 days $\times$ 25 years = 775 samples. For any given day and year, the 775-sample average difference $D$ between the ensemble-mean forecasts and the analyzed states was determined, and each ensemble member was

adjusted by subtracting the amount *D* from each member forecast. The bias correction calculations were generated separately for each location, each year, and each day. Probabilities were then calculated using the relative frequency of event occurrence from the bias-corrected ensemble.

3) "ANALOG"

The third procedure was a simplification of the analog approach described in Hamill et al. (2006) and Hamill and Whitaker (2006). The procedure for each sample location was as follows: (1) extract the ensemble-mean forecast for this sample location and Julian day/year. (2) calculate the ensemble mean of reforecasts at this location for all other years and +/- 30 days around the Julian date of the forecast in step (1) (implicitly, this step is a cross validation). (3) Determine the *n* closest reforecasts from (2) to the forecast in (1). (4) Form an ensemble from the *n* analyzed states on the dates of the closest reforecasts in step (3). (5) Determine probabilities from the relative frequency of event occurrence in the observed analogs. This procedure has the desirable property of producing an approximate conditional distribution of the observed given the forecast (Hamill and Whitaker 2006). Further, if the observed data is different in character (i.e., station data) than the forecast (i.e., gridded data), then the analog technique implicitly performs a statistical downscaling.

Probabilities were calculated for observed analogs of size $n = 20$ and 50. In subsequent plots, the skill that is plotted reflects the largest skill of the two sizes tested. In general, short-lead forecasts had the largest skill with smaller ensembles, and longer-

lead forecasts had the largest skill with larger ensembles. See also Fig. 7 from Hamill et al. (2006) for an example of this with precipitation forecasts.

c. *Forecast verification.*

1) RANK HISTOGRAMS

The rank histogram was used as a diagnostic of the ensemble's ability to reliably sample the forecast uncertainty; see Hamill (2001) for a detailed discussion of its computation and interpretation. Since the conventional application of this diagnostic assumed perfect observations, when imperfect observations or analyses are utilized, as they were here, the ensemble member forecasts should be perturbed with random noise consistent with the assumed error. Here, for Z500, T850, and T2M, random, normally distributed noise were added to each member, with an assumed standard deviation of 12.0 m, 0.6 deg C, and 1.5 deg C, respectively. The larger perturbations for T2M accounted for the additional error of representativeness (Liu and Rabier 2002) when comparing the station data to the gridded forecasts.

When tallying the rank histograms for Z500 and T850, which were on a 2.5° longitude/latitude grid, the rank of the observed relative to the sorted ensemble was weighted by the cosine of the latitude, thereby normalizing the weight applied for that sample by the grid point's area.

2) THE CONTINUOUS RANKED PROBABILITY SKILL SCORE

The continuous ranked probability score (*CRPS*; Hersbach 2000, Wilks 2006b p. 302) for a forecast on the *i*th Julian day, *j*th of *NY=26* years, and *k*th of *NP* locations is defined as

$$CRPS^{f}_{i,j,k} = \int_{-\infty}^{+\infty} \left[ F_{i,j,k}(y) - F^{o}_{i,j,k}(y) \right]^2 dy \ , \qquad (1)$$

where $F_{i,j,k}(y)$ represents the cumulative density function (cdf) as determined from one of the three forecast calibration methods discussed above, and $F^{o}_{i,j,k}(y)$ is the cdf formed from the observed datum $o_{i,j,k}$, a Heaviside function set to 0.0 for values below the observed and 1.0 for values above the observed. Note that the *CRPS* can be interpreted as the mean-absolute error of the probabilistic forecast.

We are interested in the continuous ranked probability skill score, or *CRPSS*, where the forecast skill is calculated by normalizing it by the skill of climatology. However, following Hamill and Juras (2006), we did *not* use the conventional method of calculating the *CRPSS*, commonly defined as

$$CRPSS = 1.0 - \frac{\overline{CRPS}^{f}}{\overline{CRPS}^{c}} \qquad , \qquad (2)$$

where $\overline{CRPS}^{f}$ represents the average forecast *CRPS* over all *NS* samples and $\overline{CRPS}^{c}$ represents the average *CRPS* of a forecast generated from the climatological distribution of the observations. We avoided the use eq. (2) because it was prone to overestimating skill in circumstances where the climatological distribution of the observations varied significantly from one location or one time of year to the next, which was certainly the case for these forecasts.

Instead, we performed the following procedure, similar to that proposed in Hamill and Juras (2006, eqs. 6, 8, and 9 therein), which addressed the tendency to overforecast skill. The revised *CRPSS* calculated this skill scores separately for subsets with similar observed climatologies and then arithmetically averaged the resulting *CRPSS*es.

First, assume that for all samples we have calculated the climatological uncertainty (spread) $\sigma_{i,j,k}$ . Define the mean observed climatology at this location and date as

$$\bar{o}_{i,j,k} = \sum_{p=i-15}^{p=i+15} \sum_{q=1}^{NY} \sum_{r=1}^{NP} \frac{o_{p,q,r}}{31 \times NY \times NP} \tag{3}$$

where the 31 days are centered on the Julian day *i* (in this calculation we assume no observations are missing; modification for missing observations is straightforward). Then $\sigma_{i,j,k}$ was defined as

$$\sigma_{i,j,k} = \left[ \sum_{p=i-15}^{p=i+15} \sum_{q=1}^{NY} \sum_{r=1}^{NP} \left( o_{p,q,r} - \bar{o}_{i,j,k} \right)^2 \Big/ \left( 31 \times NY \times NP - 1 \right) \right]^{1/2}. \tag{4}$$

The samples were then split up into *NC* = 8 subsets of equal size, with each subset having a distribution of $\sigma_{i,j,k}$'s that varied through a narrow range. This was achieved by sorting the $\sigma$'s from lowest to highest and dividing the ordered sample into eighths (subdivision into a larger number of subsets did not affect the skill calculation substantially). Formally, let $\mathbf{r}^S = \left[ \mathbf{r}_1^S, \ldots, \mathbf{r}_{NS/8}^S \right]$ be the associated set of sample indices (*i, j, k*) that have been placed in the *s*th of the *NC* ordered subsets. In the revised calculation of the *CRPSS*, the reference climatological score for the *s*th subset was calculated as

$$\overline{CRPS}^{c}(s) = \frac{\sum_{t=1}^{NS/8} \int_{-\infty}^{+\infty} \left[ F_{\mathbf{r}_t^S}^{C}(y) - F_{\mathbf{r}_t^S}^{O}(y) \right]^2 dy}{(NS/8)} \tag{5}$$

where $F_{\mathbf{r}_t^S}^{C}(y)$ was the climatological cdf associated with the *t*th sample in the *s*th subset. This cdf was formed from the integral of a Gaussian pdf with a mean of $\bar{o}_{\mathbf{r}_t^S}$ and a standard deviation of $\sigma_{\mathbf{r}_t^S}$. $F_{\mathbf{r}_t^S}^{O}(y)$ was the associated cdf of the observed . Equation (5) merely calculated the reference climatological *CRPS* separately for subsets with approximately equal climatological variance. Similarly, we calculated the average forecast *CRPS* for this subset as

$$\overline{CRPS}^{f}(s) = \frac{\sum_{t=1}^{NS/8} \int_{-\infty}^{+\infty} \left[ F_{\mathbf{r}_t^S}^{f}(y) - F_{\mathbf{r}_t^S}^{o}(y) \right]^2 dy}{(NS/8)} \tag{6}$$

where $F_{\mathbf{r}(t)}^{f}(y)$ was the cdf generated from the forecast method in question, raw, bias corrected, or analog. The overall *CRPSS* was then calculated as

$$CRPSS = \frac{1}{NC} \sum_{s=1}^{NC} \left( 1 - \frac{\overline{CRPS}^{f}(s)}{\overline{CRPS}^{C}(s)} \right). \tag{7}$$

A disadvantage of eq. (7) is that it is less resistant to outliers (Wilks 2006b, p. 23) than the conventional method of eq. (2). Since skill is bounded above by 1.0 but unbounded below, if subsets have particularly low skill, they may strongly depress the overall skill.

For 500 hPa height and 850 hPa temperature, eqs. (5) and (6) were modified slightly to account for the different areas encompassed by different grid points, weighting the samples the cosine of their latitude.

## 3. **Results**

### *a. Rank histograms*

Figures 2-4 show forecast rank histograms at 1, 4, and 7-days lead. The shaded bars denote the rank populations of the raw forecasts, while the solid lines indicate the distribution of rank populations after the bias correction. At day 1, all raw forecasts underforecasted the variable in question to varying degrees, reflected in the greater population of the higher ranks. After bias correction, the distributions were more evenly centered, but Z500 had too little population of the extreme ranks while the T2M exhibited way too much population of the extreme ranks. After bias correction, T850 exhibited a slight overpopulation of the extreme ranks.

At day 4, the same underforecasting bias was evident in the raw forecasts, but after the bias correction, now all the forecasts exhibited some overpopulation of the extreme ranks, with T2M again exhibiting the largest overpopulation. The characteristics at day 7 were similar.

Figure 5 provides further illustration of the larger systematic errors of T2M. Here we have plotted, averaged over all samples, the absolute value of the bias divided by the observational climatological uncertainty (the standard deviation of the observed about its climatology, as in eq. 4). It's readily apparent that at short leads, the bias in Z500 was a small fraction of the climatological spread, while it was a large fraction for T2M, and in

between for T850. As forecast lead increased, the ratio of bias to uncertainty increased for Z500 and T850, indicating that the bias was increasing to more like the high levels seen with T2M.

Judging from the rank histograms and magnitude of bias, one would expect that a bias correction of T2M forecasts would be essential at all leads, while bias correction for Z500 would become increasingly necessary at longer leads. However, since the T2M rank histograms remained non-uniform even after bias correction, this suggests that a technique like the analog that adjusted probabilities for the tendency of the observed to lie outside the range of the ensemble should more dramatically improve the skill of T2M forecasts relative to T850 or Z500. We now test this hypothesis.

*b. CRPSS*

Figure 6 shows the *CRPSS* of Z500 as a function of forecast lead. Correction for gross bias substantially improved the skill relative to the raw forecasts, though the application of the analog method, which also addressed biases in the spread, produced little additional benefit during the first week of the forecast. During the second week, it produced a modest improvement in forecast skill.

Figure 7 shows the *CRPSS* of T850 forecasts. In comparison to Z500, the raw forecasts lost skill more quickly, and there was still a substantial impact from the correction of gross bias. However, after the first few days of the forecast, there was a more substantial increase in forecast skill from applying the analog technique and its implicit spread plus bias corrections relative to the skill increase for Z500. A 4.5-day

analog T850 forecast had skill similar to an ~3-day bias-corrected forecast or a 1.5-day raw forecast.

Figure 8 shows the *CRPSS* of T2M forecasts. The raw forecasts, even on the first day, had skill worse than the reference climatology, though the bias correction boosted the skill to above zero. Relative to bias-corrected T850 and Z500, the analog method added more skill, with a 3-day analog forecast having approximately the same skill as a 1-day bias-corrected forecast. This dramatic improvement conformed with our expectations given the shape of the rank histograms, previously discussed.

Why were the raw forecasts so unskillful? Figure 9 shows the average *CRPS* (the raw score, not the skill score) of the forecasts calculated separately for the lowest and highest of the eight climatological uncertainty subsets used in eq. (7). The subset with large climatological uncertainty showed a raw *CRPS* smaller than the climatological *CRPS* at early leads but larger than the climatological *CRPS* at longer leads. For the subset with small climatological uncertainty, the average *CRPS* of the forecast was uniformly larger than the *CRPS* of climatology, indicating that at day 1 the errors had already saturated. Hence, the reason that the overall T2M *CRPSS* was so small was because there were many samples where the *CRPSS* of climatology was small enough to be an extremely tough reference forecast to beat. Consequently, the overall skill was diminished substantially by the negative skill at for these samples. Note also in Fig. 9 that after the application of the analog technique, the forecast *CRPS* was reduced to less than the *CRPS* of climatology for both large and small uncertainty subsets.

## 4. **Conclusions.**

An examination of the effects of calibration was performed for 3 variables, 500-hPa geopotential height, 850-hPa temperature, and 2-meter temperature forecasts. Two calibration methods were examined, a correction for a gross bias in the forecast and an analog method that implicitly corrected for both bias, spread, and downscaling. 500-hPa geopotential height was the variable with the most skill before calibration and the least dramatically improved by calibration. A simple bias correction achieved most of the improvement in skill. For 850-hPa temperature, the raw forecast skill was somewhat diminished, and there was a more substantial additional increase in skill relative to the bias correction when the analog technique was applied. For 2-m temperature forecasts, the raw forecasts were unskillful, and the application of the analog technique produced a dramatic increase in skill relative to the simple bias correction.

Of course, 2-m temperature forecasts will have a wider variety of users than 850-hPa temperature or 500-hPa geopotential. Previously, we have demonstrated with precipitation calibration experiments (e.g., Hamill et al. 2004, 2006, Hamill and Whitaker 2006) that precipitation forecasts were particularly difficult to calibrate and required a large forecast sample to improve the forecasts substantially. Here, surface-temperatures turned out to be the most difficult to forecast correctly without calibration and the most amenable to forecast improvements through calibration. Previous work (Wilks and Hamill 2006) showed that large training samples were necessary to fully realize the gains from statistical calibration. Taken together, these results reinforce the our previous assertion that large reforecast data sets may be particularly helpful in the improvement of probabilistic forecasts of the variables that are most directly relevant to forecast users.

We again recommend that the computation of reforecasts and the subsequent calibration of the forecasts become regular parts of the numerical weather prediction process.

ACKNOWLEDGMENTS

We gratefully acknowledge the programming support of our former colleague, Xue Wei, who recently left to return to China. We appreciate the fruitful discussions with the NCEP ensemble group including Zoltan Toth, Bo Cui, Dingchen Hou, and Yuejian Zhu.

REFERENCES

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132-1147.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman. 2005: Calibrated probabilistic forecasting using ensemble Model Output Statistics and minimum CRPS estimation. *Mon. Wea. Rev*., **133**, 1098–1118.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev*., **126**, 711-724.

----------, 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550-560.

----------, J. S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev*., **132**, 1434-1447.

----------, ----------, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc*., **87**, 33-46.

----------, and ----------, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev*., **134**, 3209-3229.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc*., in press. Available at http://www.cdc.noaa.gov/people/tom.hamill/skill_overforecast_QJ_v2.pdf .

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559-570.

Kalnay, E., and co-authors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437-472.

Liu, Z.-Q. and F. Rabier, 2002: The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study. *Quart. J. Roy. Meteor. Soc.*, **128**, 1367–1386.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155-1174.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16-30.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.* **74**, 2317-2330.

----------, and ---------- , 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Royal Meteor. Soc.*, **131**, 965-986.

Wilks, D. S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Apps.*, **13**, 243-256.

----------, 2006b: *Statistical Methods in the Atmospheric Sciences (2nd Ed)*. Academic Press. 627 pp.

----------, and T. M. Hamill, 2006: Comparison of ensemble-MOS methods using GFS reforecasts. Mon. Wea. Rev., in press. Available at http://www.cdc.noaa.gov/people/tom.hamill/WilksHamill_emos.pdf .

FIGURE CAPTIONS

**Figure 1**: Station locations for calibration of 0000 UTC 2-m temperature forecasts.

**Figure 2**: Rank histograms for 1-day forecasts. (a) 500-hPa geopotential height, (b) 850-hPa temperature, and (c) 2-m temperature. Shaded bars denote the rank histogram before bias correction, and dark lines after the bias correction.

**Figure 3**: As in Fig. 2, but for 4-day forecasts.

**Figure 4**: As in Fig. 2, but for 7-day forecasts.

**Figure 5**: Ratio of gross bias to the climatological uncertainty as a function of time of year and forecast lead. (a) Z500, (b) T850, and (c) T2M.

**Figure 6**: Northern-Hemispheric average CRPSS as a function of forecast lead and calibration method.

**Figure 7**: As in Fig. 5, but for 850 hPa temperature.

**Figure 8**: As in Fig. 5, but for T2M forecasts.

**Figure 9**: Average CRPS of raw and climatological forecasts for (a) the $1/8^{th}$ subset of samples with the highest climatological uncertainty, and (b) the $1/8^{th}$ subset with the lowest climatological uncertainty.

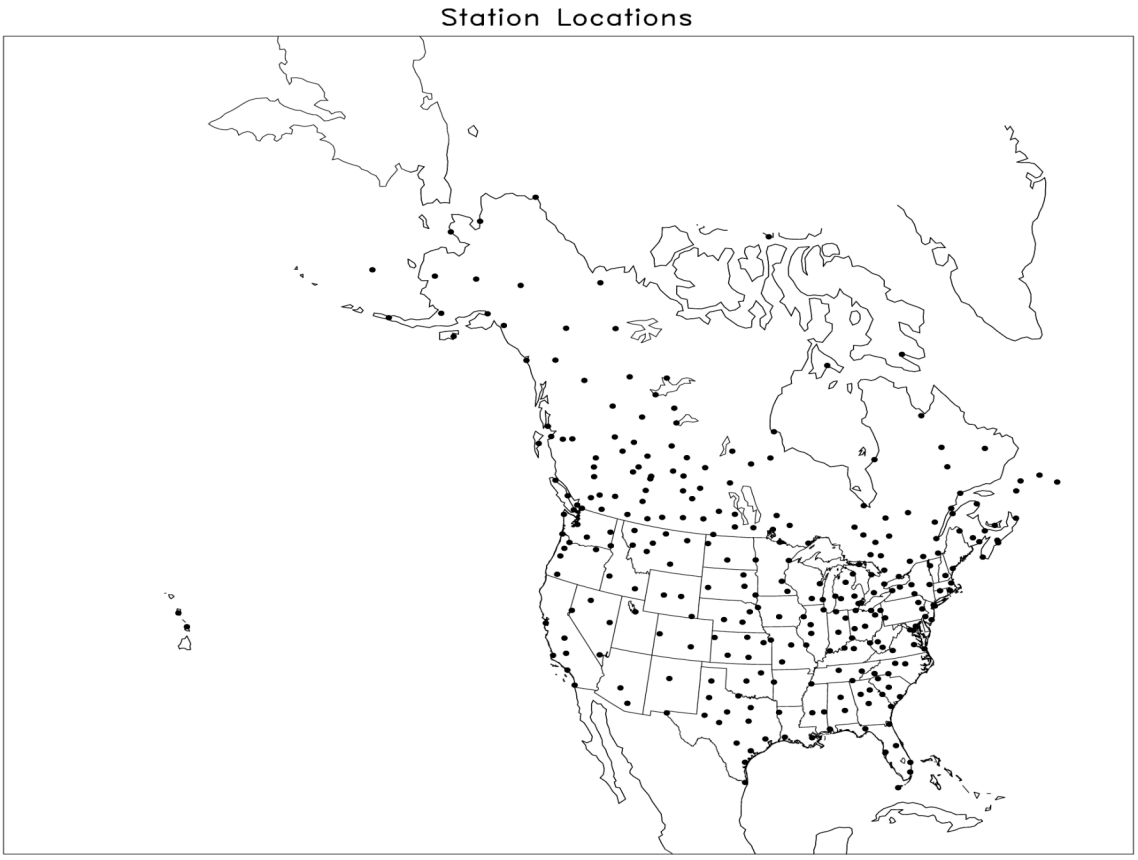


**Figure 1**: Station locations for calibration of 0000 UTC 2-m temperature forecasts.

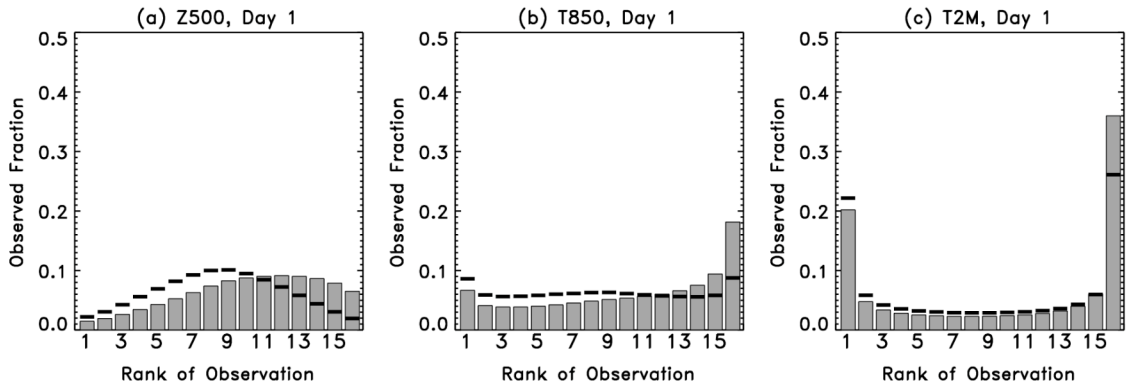


**Figure 2**: Rank histograms for 1-day forecasts. (a) 500-hPa geopotential height, (b) 850-hPa temperature, and (c) 2-m temperature. Shaded bars denote the rank histogram before bias correction, and dark lines after the bias correction.

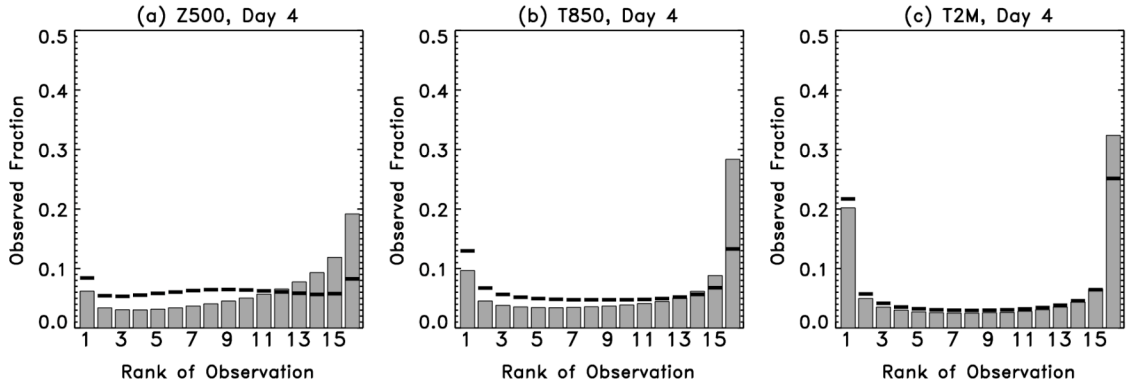


**Figure 3**: As in Fig. 2, but for 4-day forecasts.

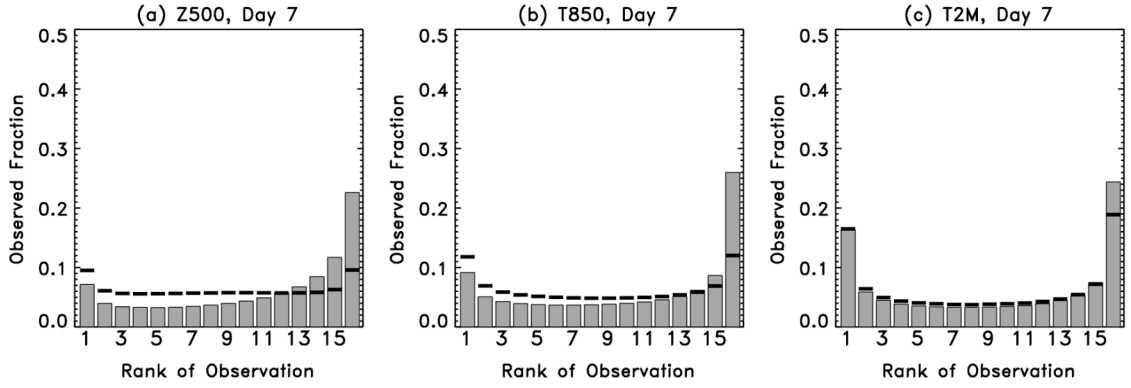


**Figure 4**: As in Fig. 2, but for 7-day forecasts.

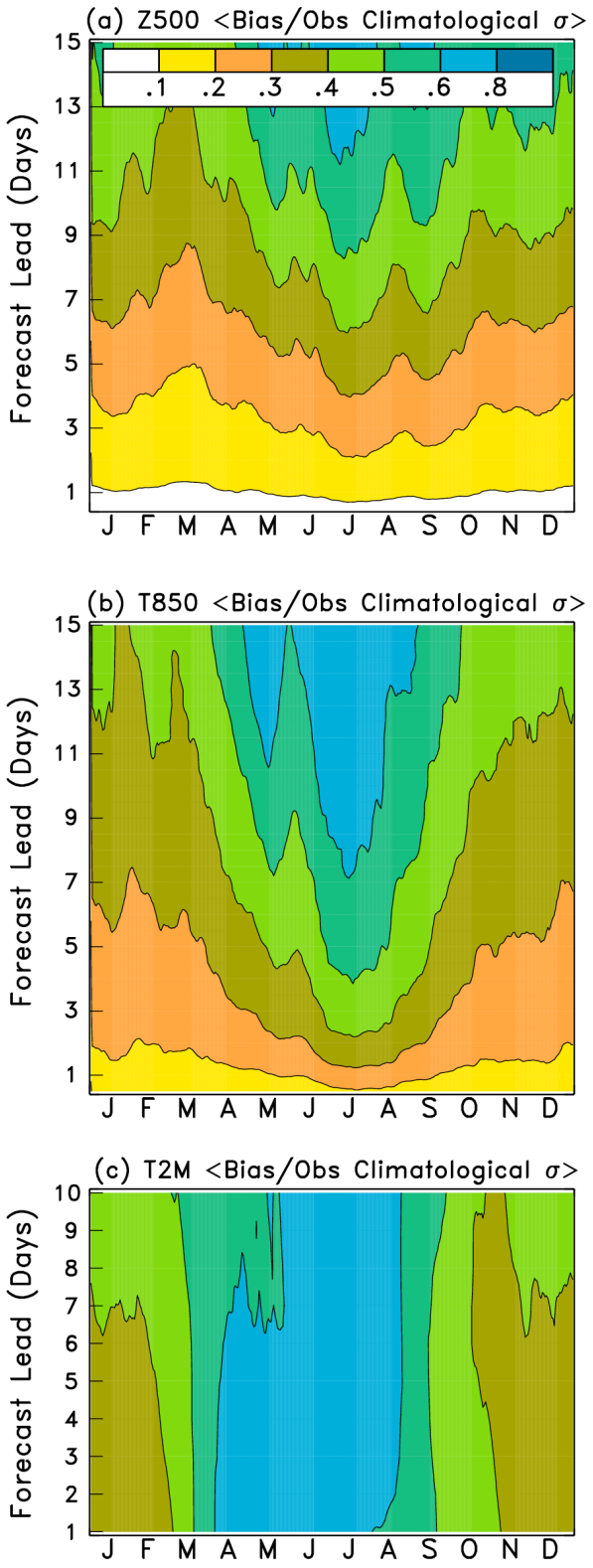


**Figure 5**: Ratio of gross bias to the climatological uncertainty as a function of time of year and forecast lead. (a) Z500, (b) T850, and (c) T2M.

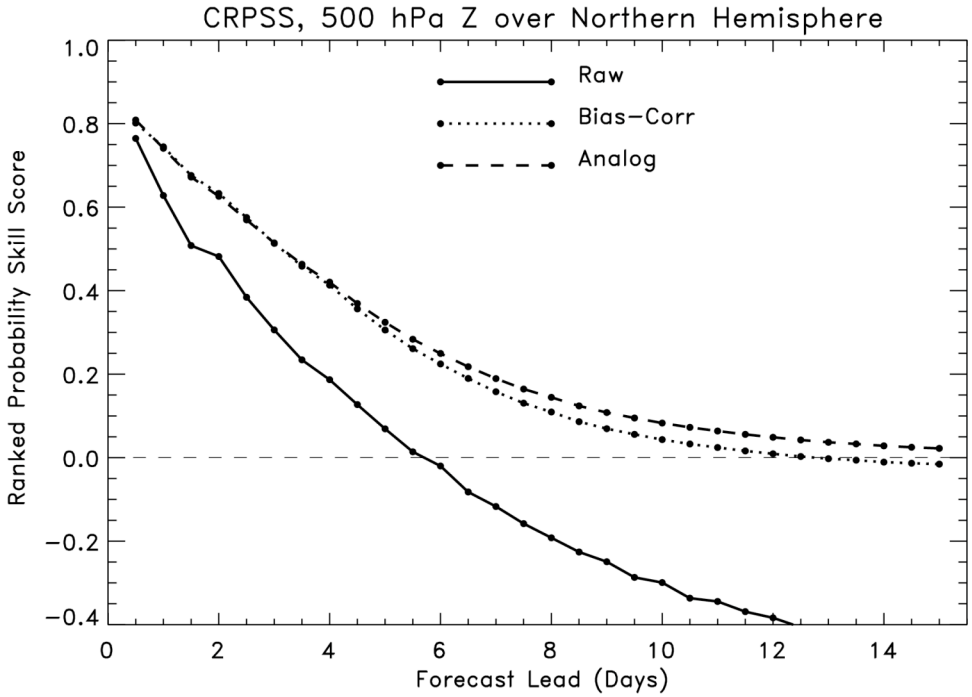


**Figure 6**: Northern-Hemispheric average CRPSS as a function of forecast lead and calibration method.

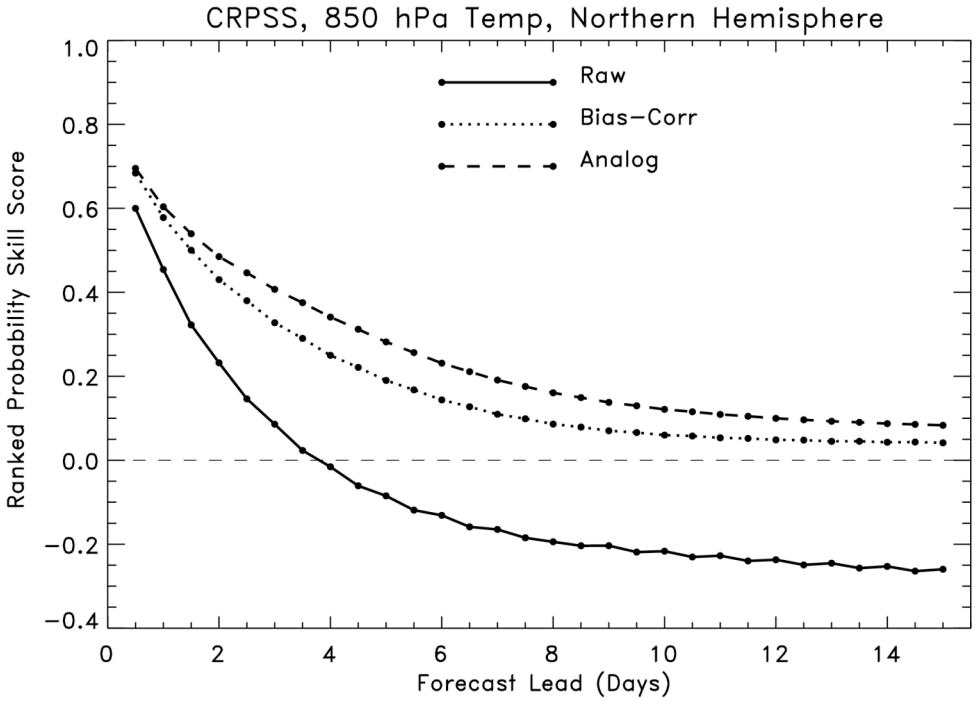


**Figure 7**: As in Fig. 5, but for 850 hPa temperature.

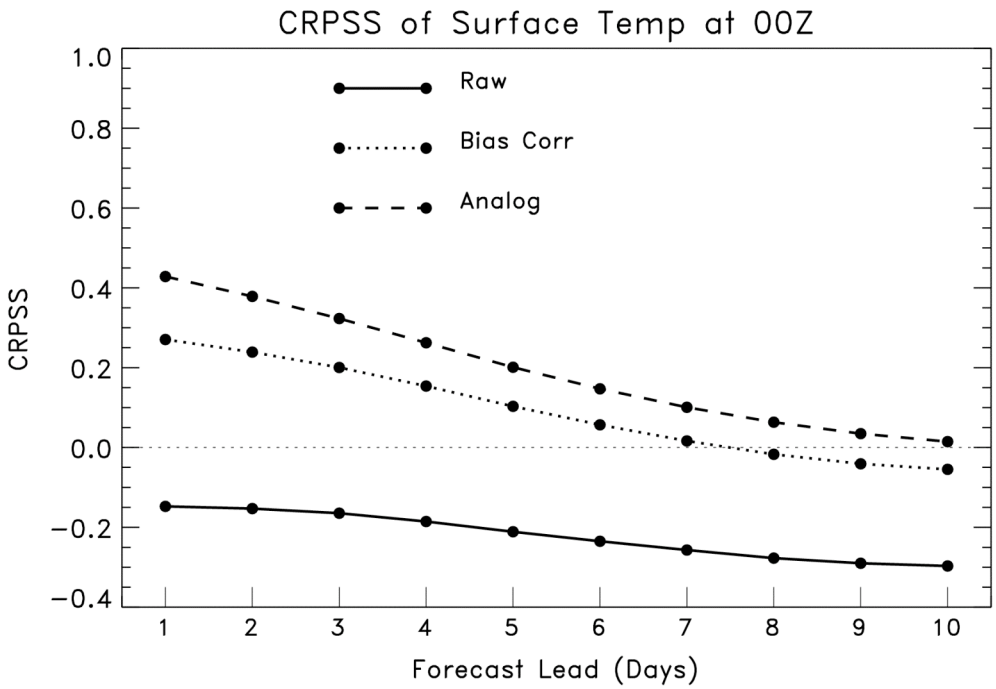


**Figure 8**: As in Fig. 5, but for T2M forecasts.

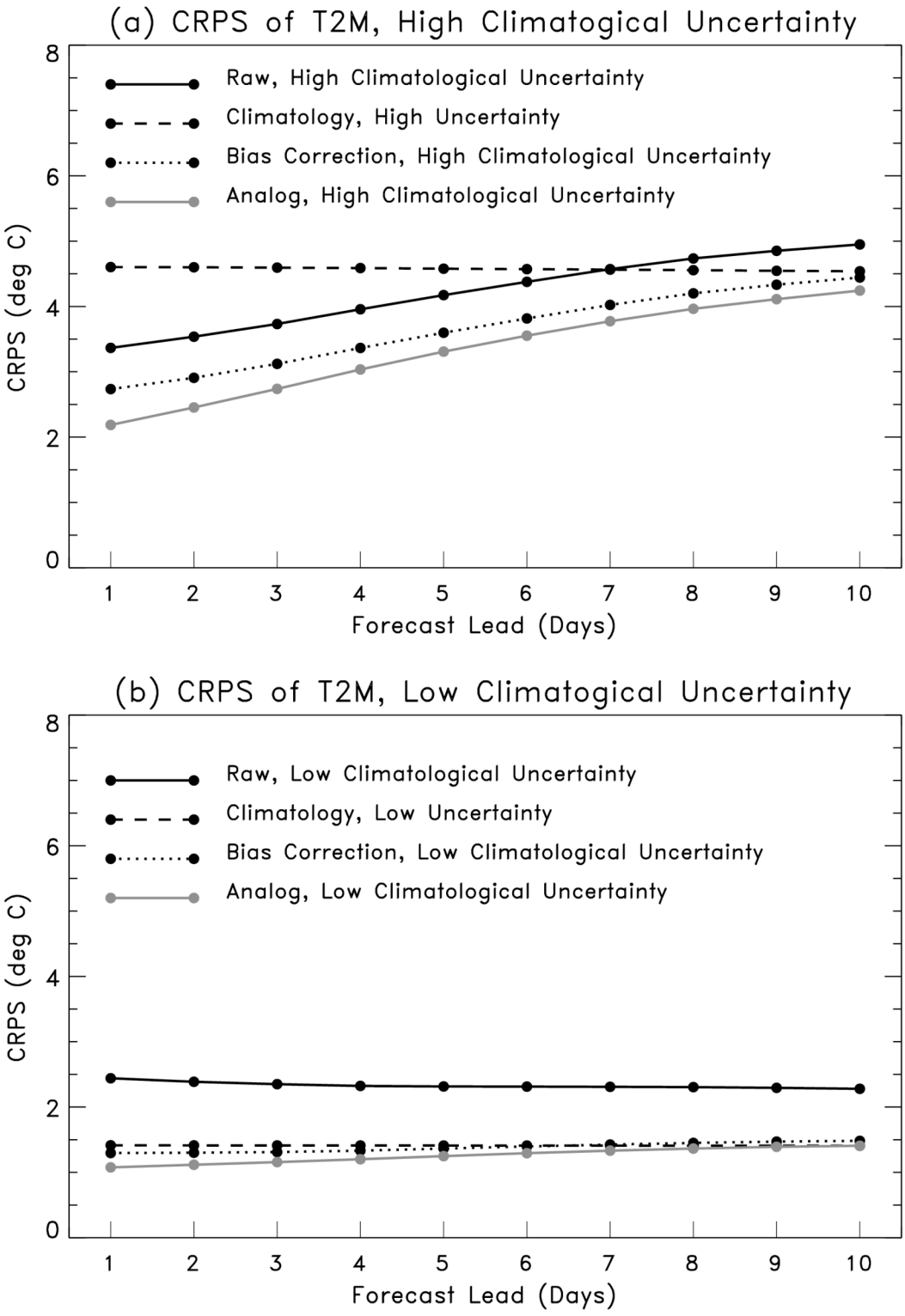


**Figure 9**: Average *CRPS* of raw, bias-corrected, analogs, and climatological forecasts for (a) the 1/8$^{th}$ subset of samples with the highest climatological uncertainty, and (b) the 1/8$^{th}$ subset with the lowest climatological uncertainty.